

What is classification?

- A supervised learning approach
- Categorizing some unknown items into a discrete set of categories or “classes”
- The target attribute is a categorical variable

How does classification work?

Classification determines the class label for an unlabeled test case.

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Categorical Variable

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	0

Prediction



Predicted Labels

Example of multi-class classification

Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Categorical Variable

Modeling

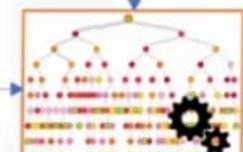
Prediction

Classifier

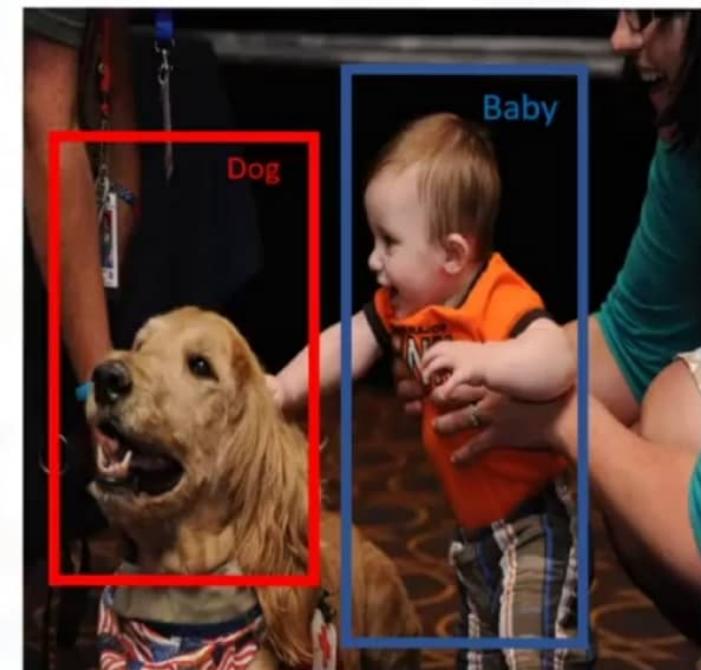
Age	Sex	BP	Cholesterol	Na	K	Drug
36	F	LOW	HIGH	0.697	0.069	DrugX

DrugX

Predicted Labels



Classification applications



Classification algorithms in machine learning

- Decision Trees (ID3, C4.5, C5.0)
- Naïve Bayes
- Linear Discriminant Analysis
- k -Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM) 

Intro to KNN

X: Independent variable Y: Dependent variable

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value Label

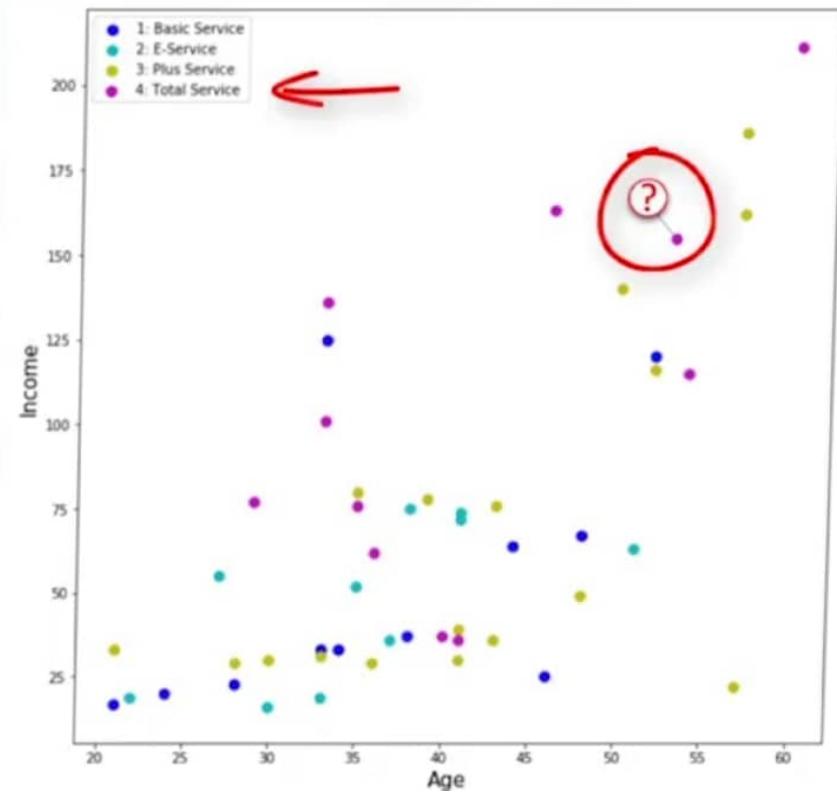
Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Determining the class using 1st KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

1-NN

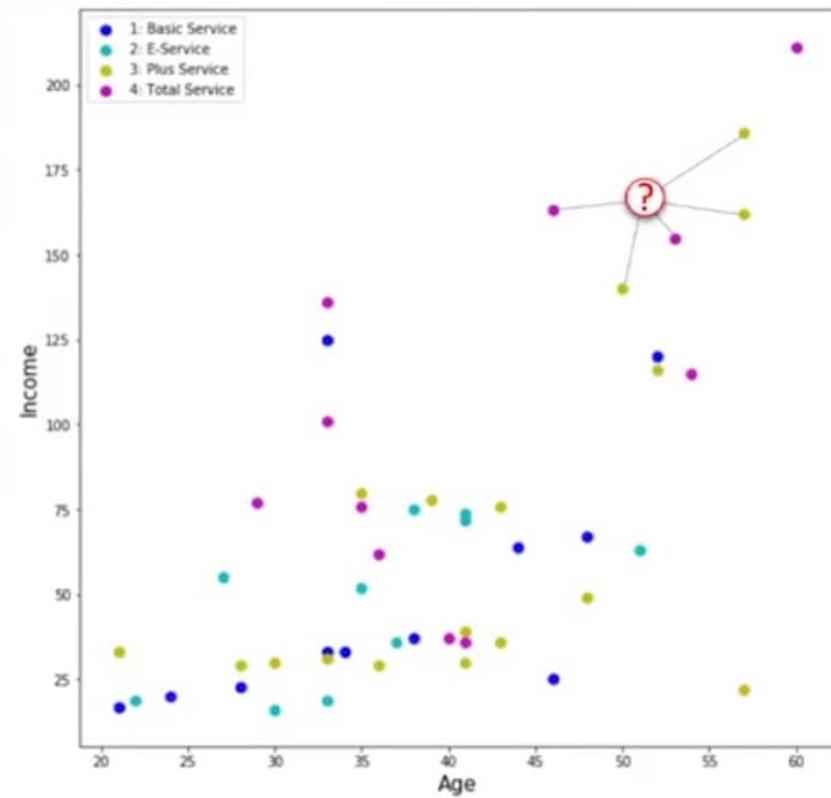
→ 4: Total Service



Determining the class using the 5 KNNs

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

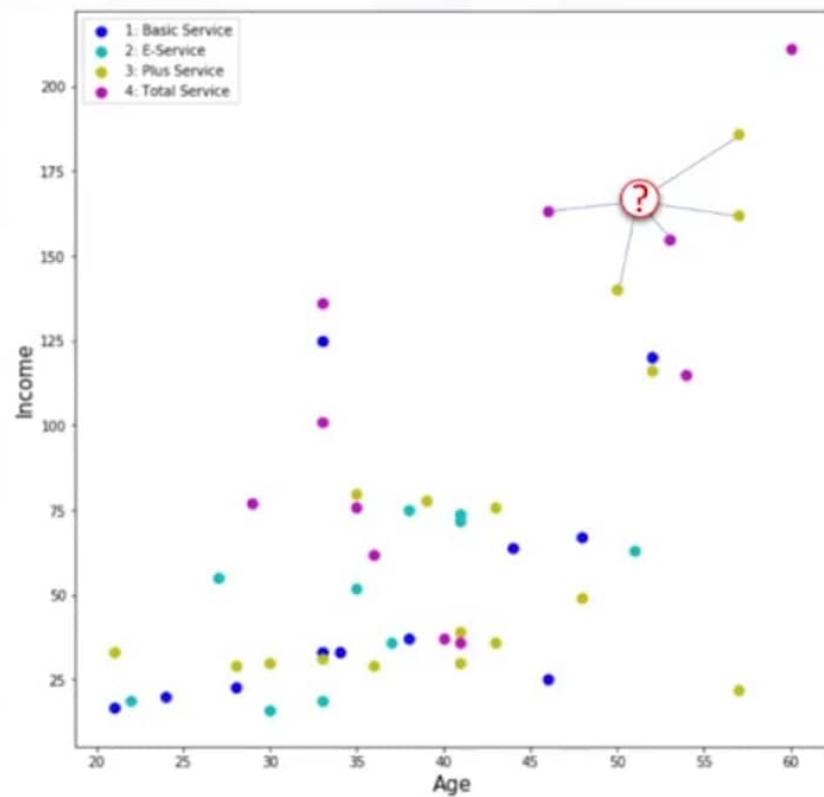
5-NN → 3: Plus Service





What is K-Nearest Neighbor (or KNN)?

- A method for **classifying** cases based on their similarity to other cases
- Cases that are near each other are said to be “**neighbors**”
- Based on **similar cases with same class labels** are near each other



The K-Nearest Neighbors algorithm

1. Pick a value for K.
2. Calculate the distance of unknown case from all cases.
3. Select the K-observations in the training data that are “nearest” to the unknown data point.
4. Predict the response of the unknown data point using the most popular response value from the K-nearest neighbors.

Calculating the similarity/distance in a 1-dimensional space



Customer 1

Age

34



Customer 2

Age

30

$$\text{Dis } (x_1, x_2) = \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2}$$

$$\text{Dis } (x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

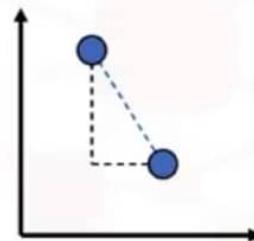
Calculating the similarity/distance in a 2-dimensional space



Customer 1	
Age	Income
34	190



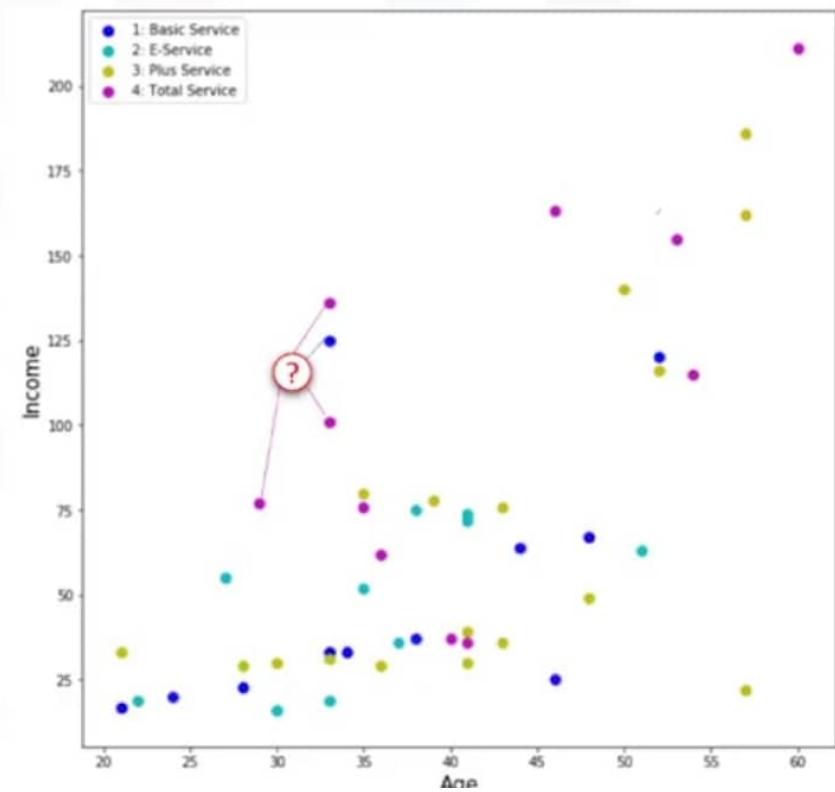
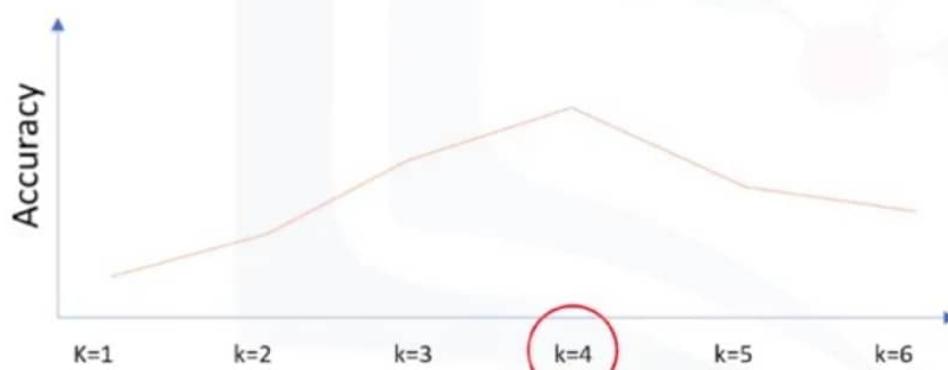
Customer 2	
Age	Income
30	200



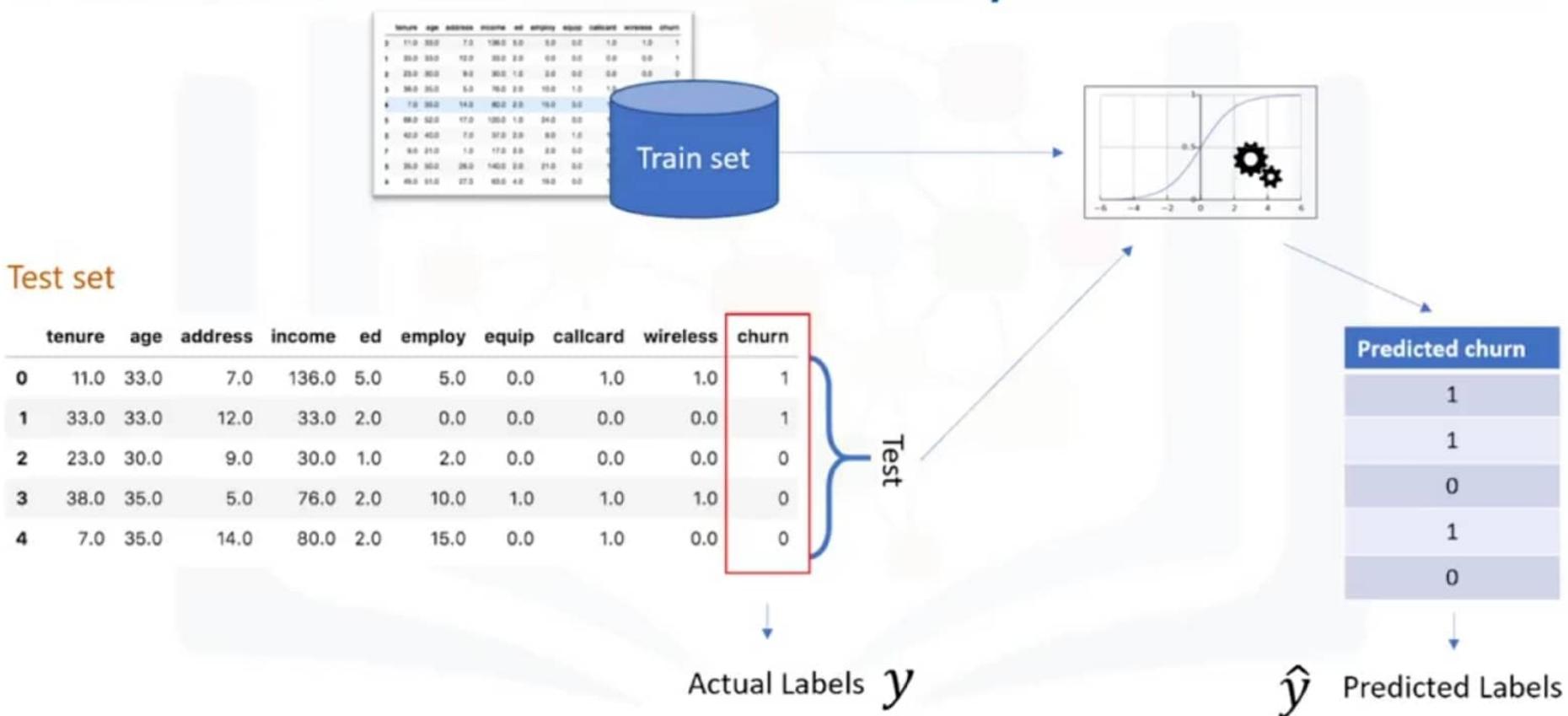
$$\begin{aligned}\text{Dis}(x_1, x_2) &= \sqrt{\sum_{i=0}^n (x_{1i} - x_{2i})^2} \\ &= \sqrt{(34 - 30)^2 + (190 - 200)^2} = 10.77\end{aligned}$$

What is the best value of K for KNN?

- K =1 class 1
- K =20 ?



Classification accuracy



Jaccard index

y : Actual labels

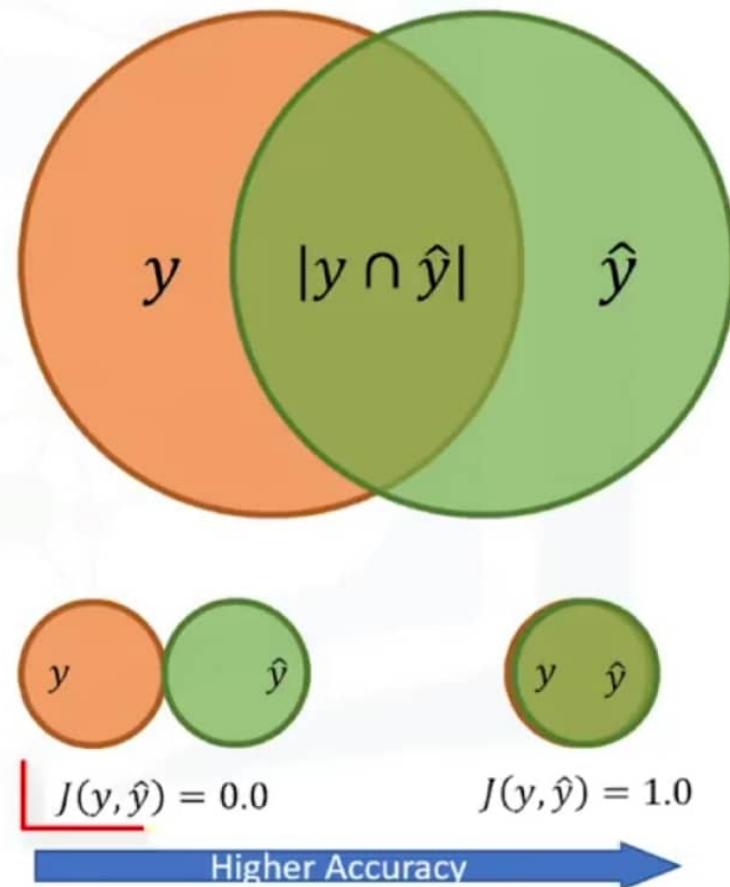
\hat{y} : Predicted labels

$$J(y, \hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

$y: [0, 0, 0, 0, 0, 1, 1, 1, ,1,1]$

$\hat{y}: [1, 1, 0, 0, 0, 1, 1, 1, ,1,1]$

$$J(y, \hat{y}) = \frac{8}{10+10-8} = 0.66$$



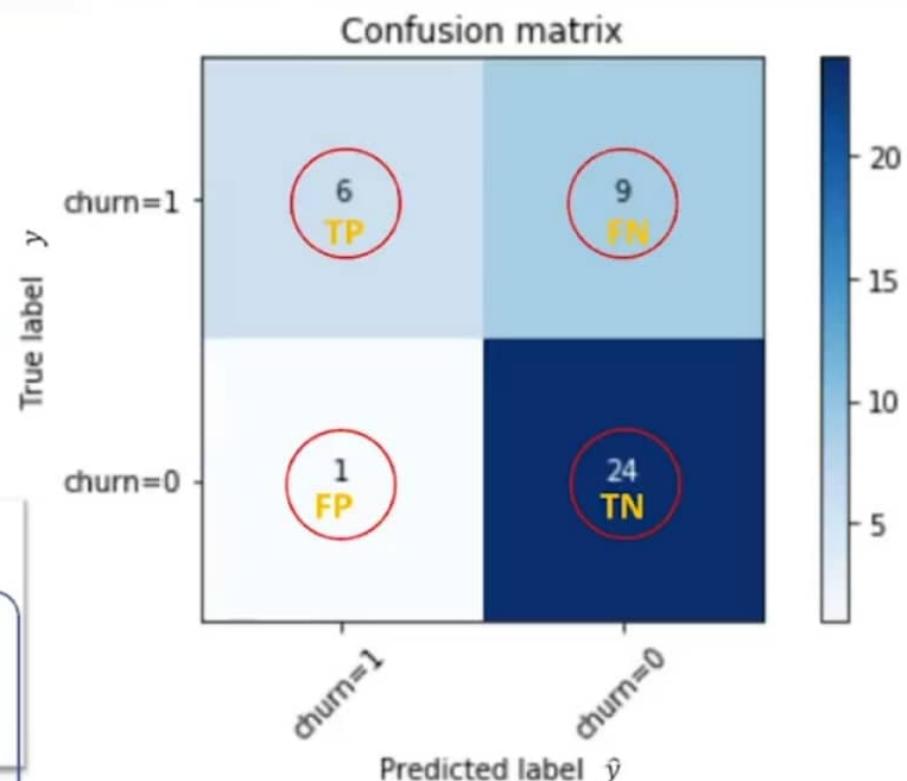
F1-score

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1-score = $2x (prc \times rec) / (prc+rec)$

F1-score: [0.00] ... 0.20 0.55 0.83 ... [1.00]
Higher Accuracy →

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55

Avg Accuracy = 0.72

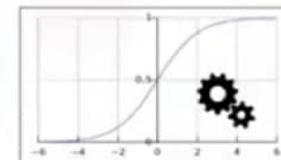


Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.

Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Actual Labels y

Predicted churn	LogLoss
0.91	0.11
0.13	2.04
0.04	0.04
0.23	0.26
0.43	0.56

$$\text{LogLoss} = 0.60$$

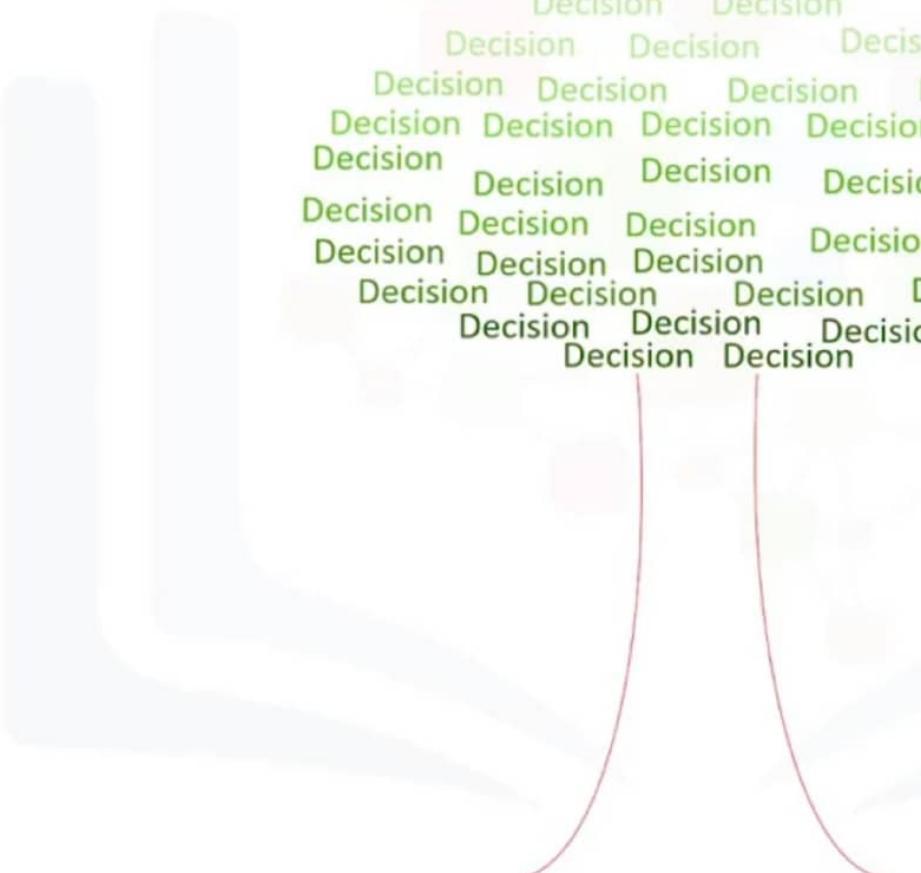
\hat{y} Predicted Probability

$$\text{LogLoss} = -\frac{1}{n} \sum (y \times \log(\hat{y}) + (1 - y) \times \log(1 - \hat{y}))$$

LogLoss: 0.00 ... 0.35 0.60 ... 1.00

Higher Accuracy

What is a decision tree?



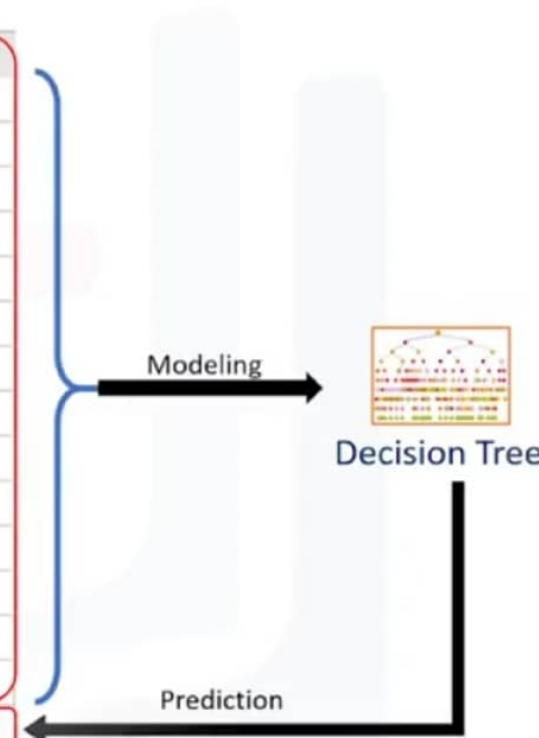
```
Decision Decision  
Decision Decision Decision  
Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision  
Decision Decision Decision Decision Decision
```

The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree.

[Narendra Nath Joshi](#)

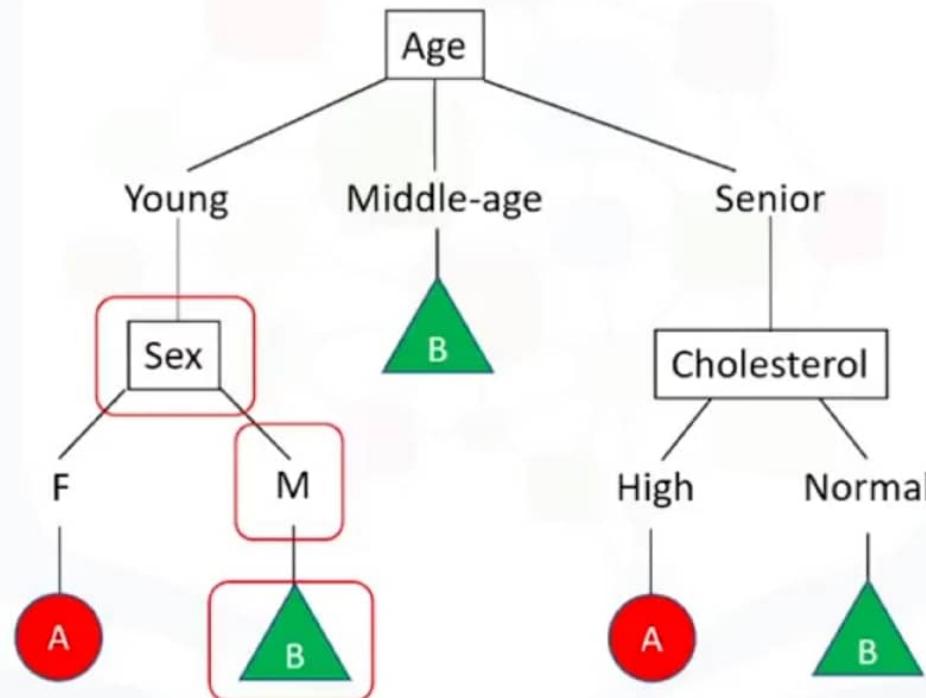
How to build a decision tree?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?



Building a decision tree with the training set

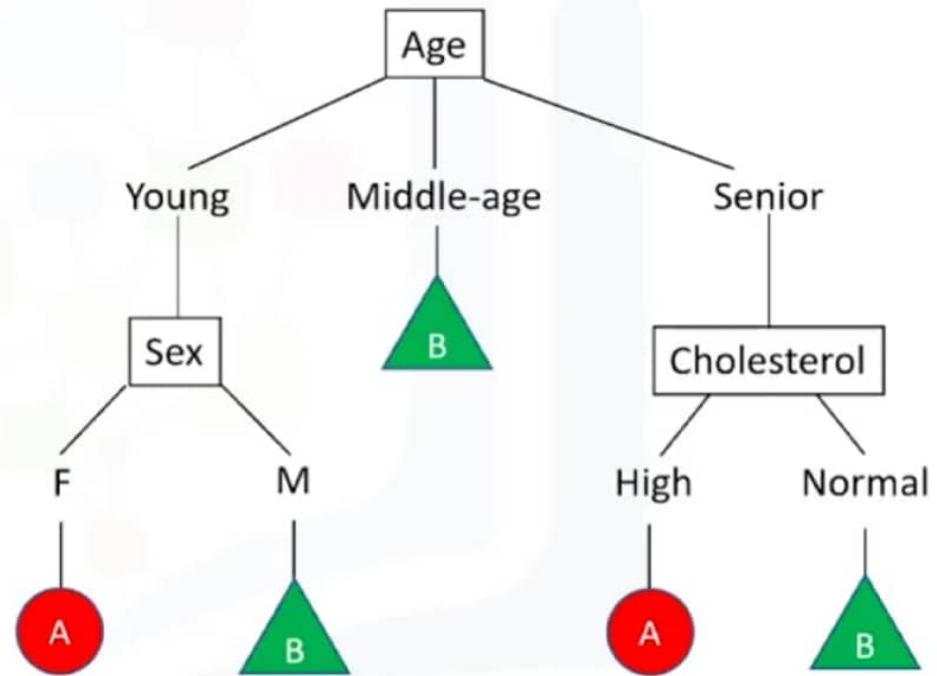
Drug B
Drug A



- Each **internal node** corresponds to a test
- Each **branch** corresponds to a result of the test
- Each **leaf node** assigns a classification

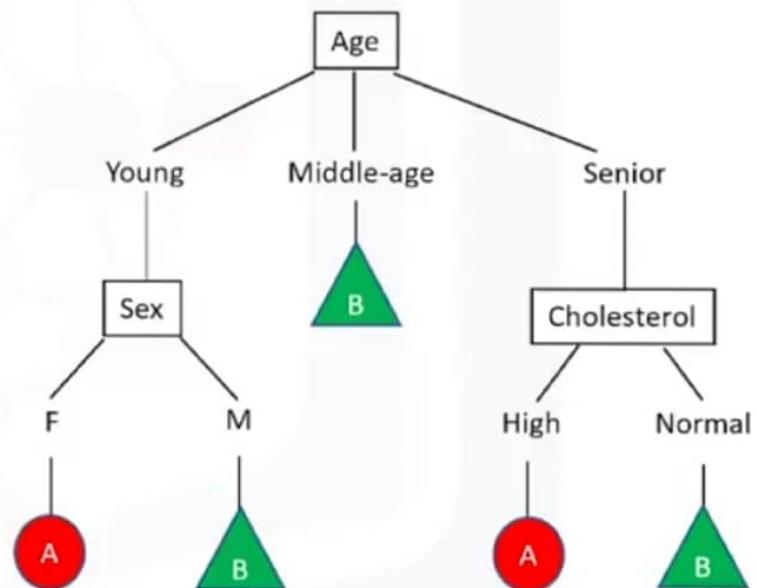
Decision tree learning algorithm

1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to step 1.



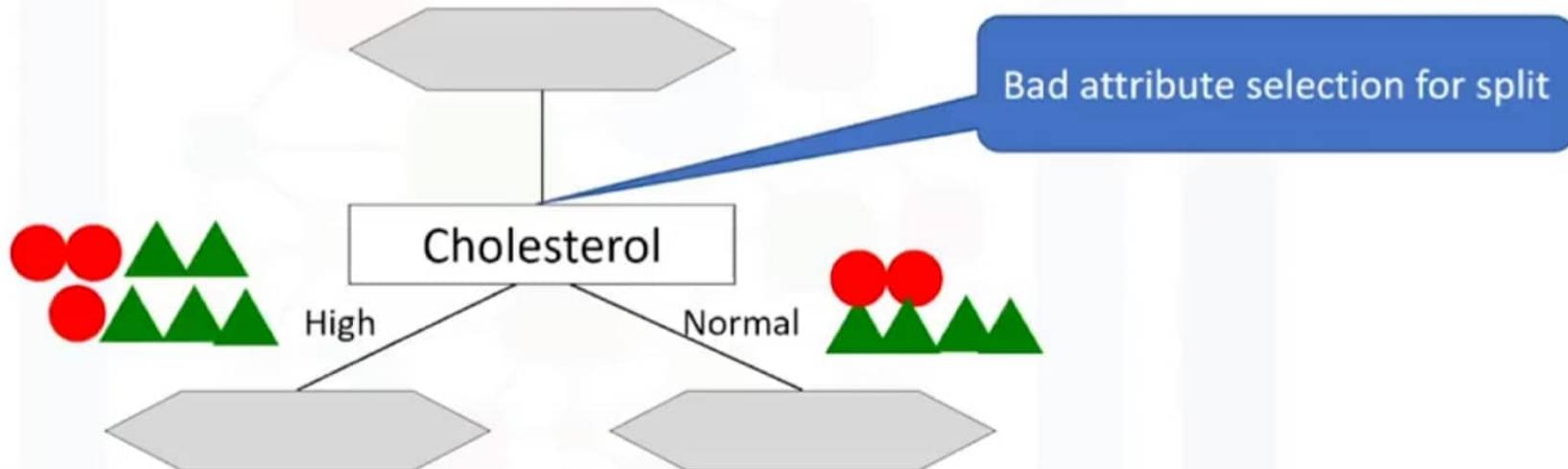
How do you build a decision tree?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?



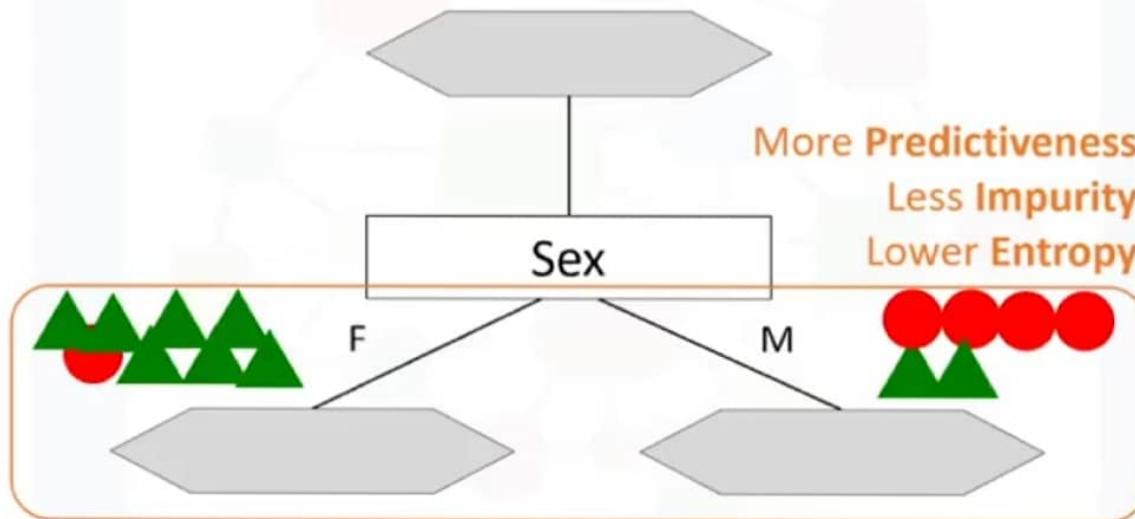
Which attribute is the best ?

Drug B
Drug A



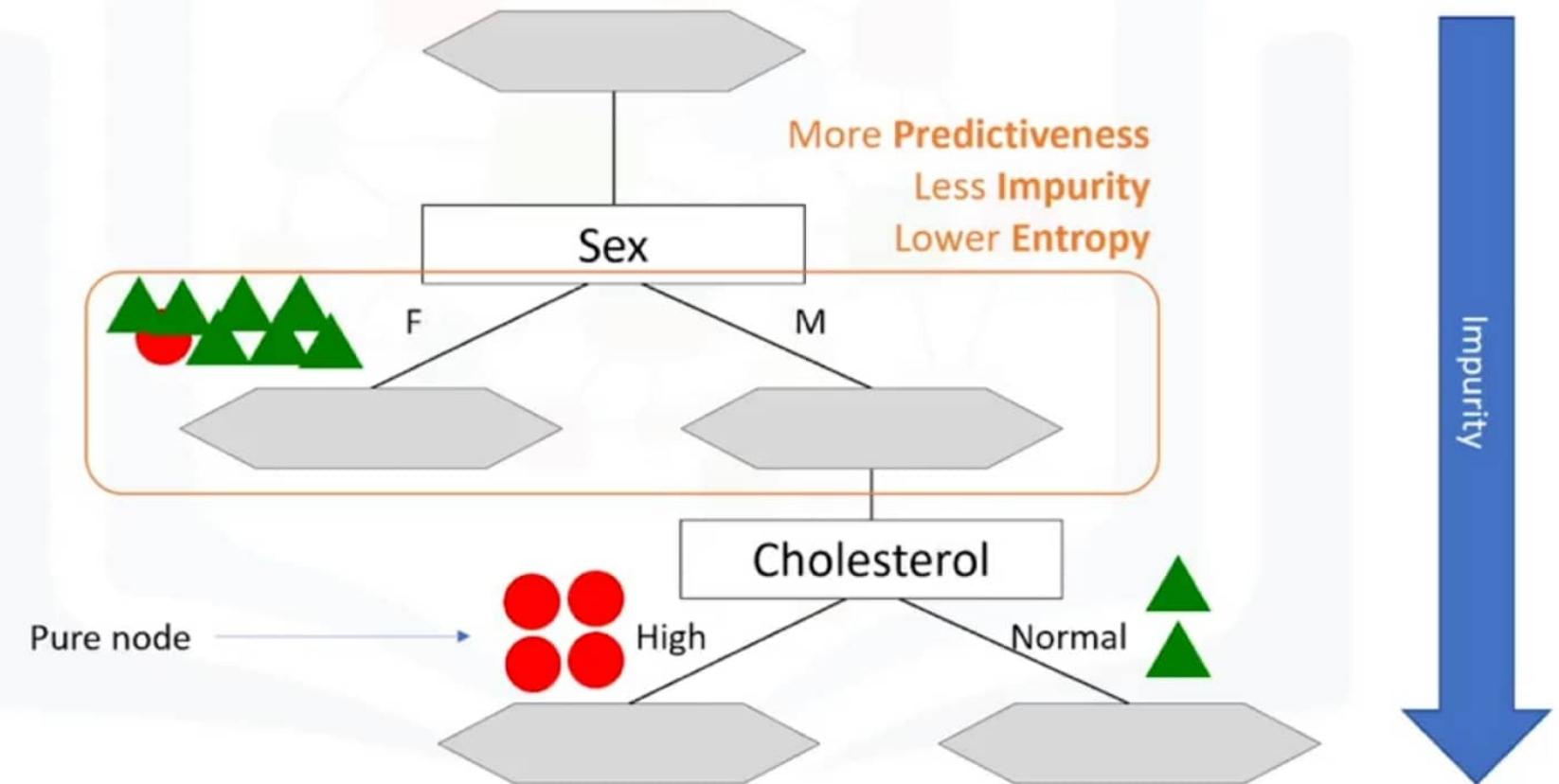
Which attribute is the best ?

Drug B
Drug A



Which attribute is the best ?

▲ Drug B
● Drug A

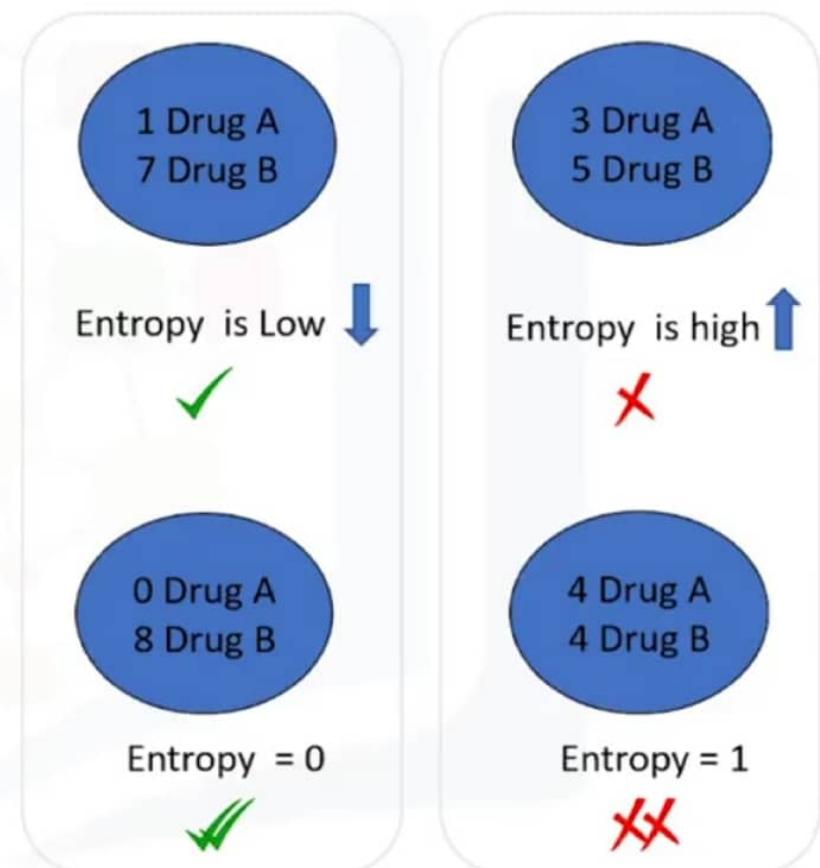


Entropy

- Measure of randomness or uncertainty

$$\text{Entropy} = - p(A)\log(p(A)) - p(B)\log(p(B))$$

The lower the Entropy, the less uniform the distribution, the purer the node.

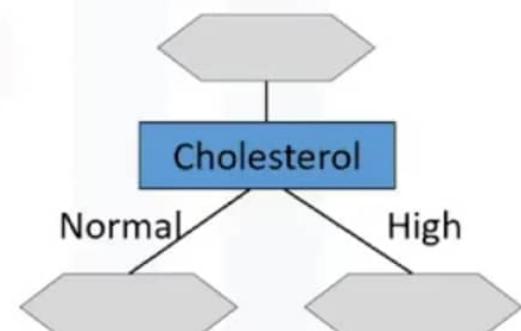


Is 'Cholesterol' the best attribute?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

E = 0.940



S: [6 B, 2 A]

E = 0.811

S: [3 B, 3 A]

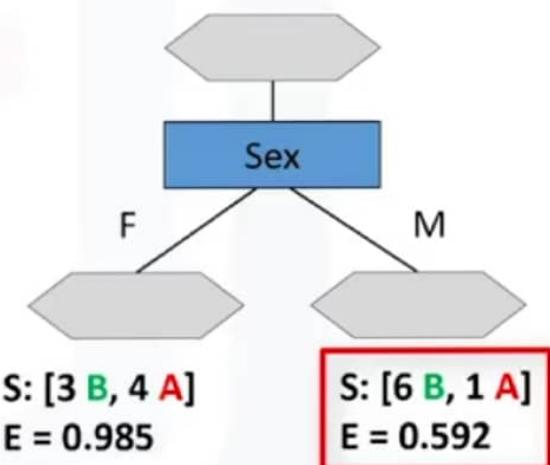
E = 1.00

What about 'Sex'?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

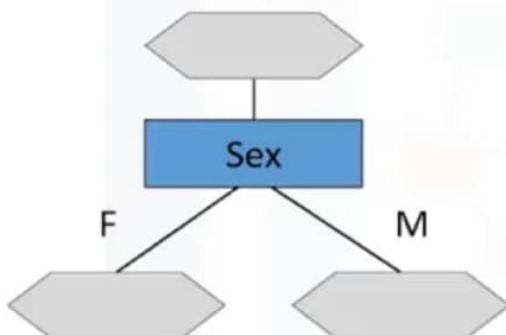
E = 0.940



Which attribute is the best?

S: [9 B, 5 A]

E = 0.940



S: [3 B, 4 A]

E = 0.985

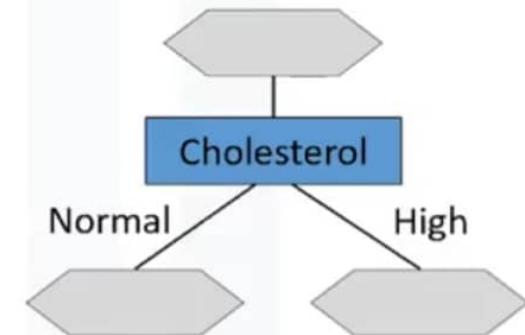
S: [6 B, 1 A]

E = 0.592

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

E = 0.940



S: [6 B, 2 A]

E = 0.811

S: [3 B, 3 A]

E = 1.00

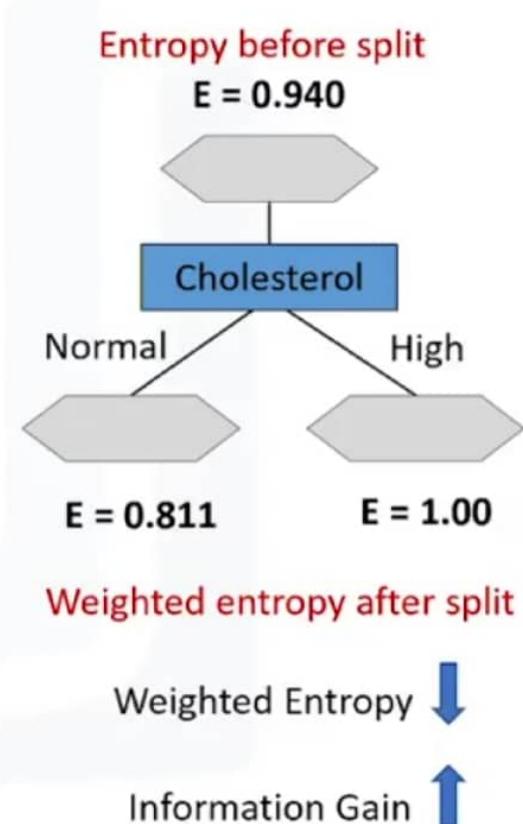
?

The tree with the higher Information Gain after splitting.

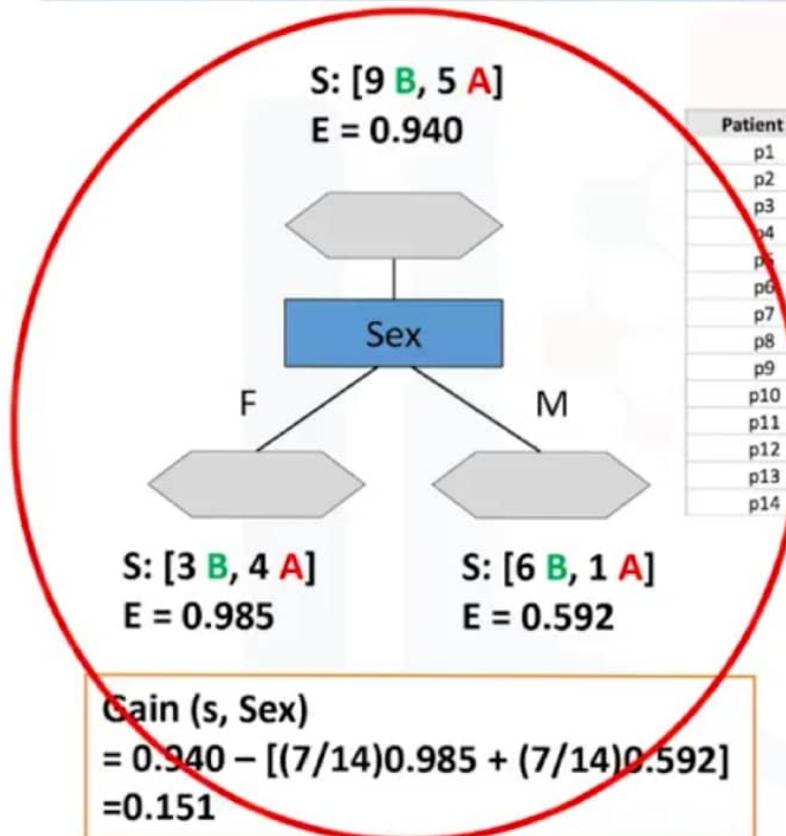
What is information gain?

Information gain is the information that can increase the level of certainty after splitting.

$$\text{Information Gain} = (\text{Entropy before split}) - (\text{weighted entropy after split})$$

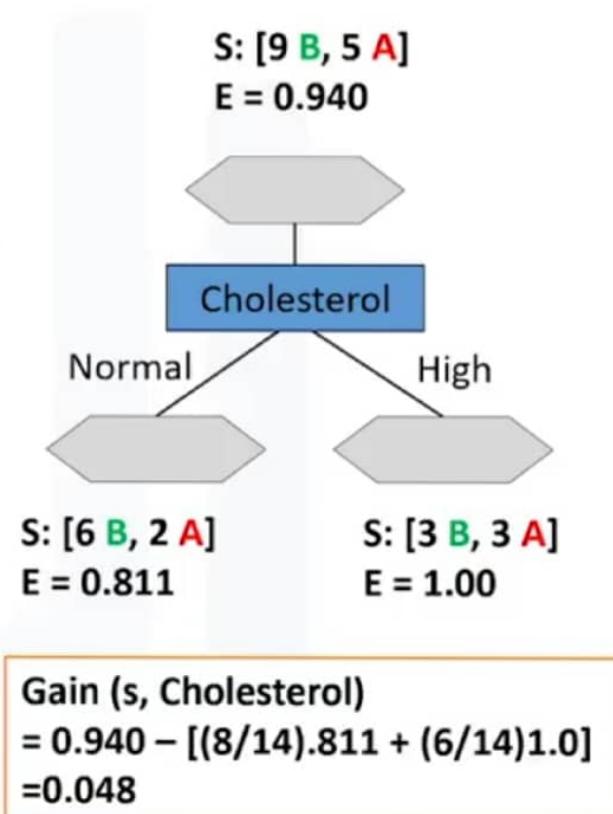


Which attribute is the best?

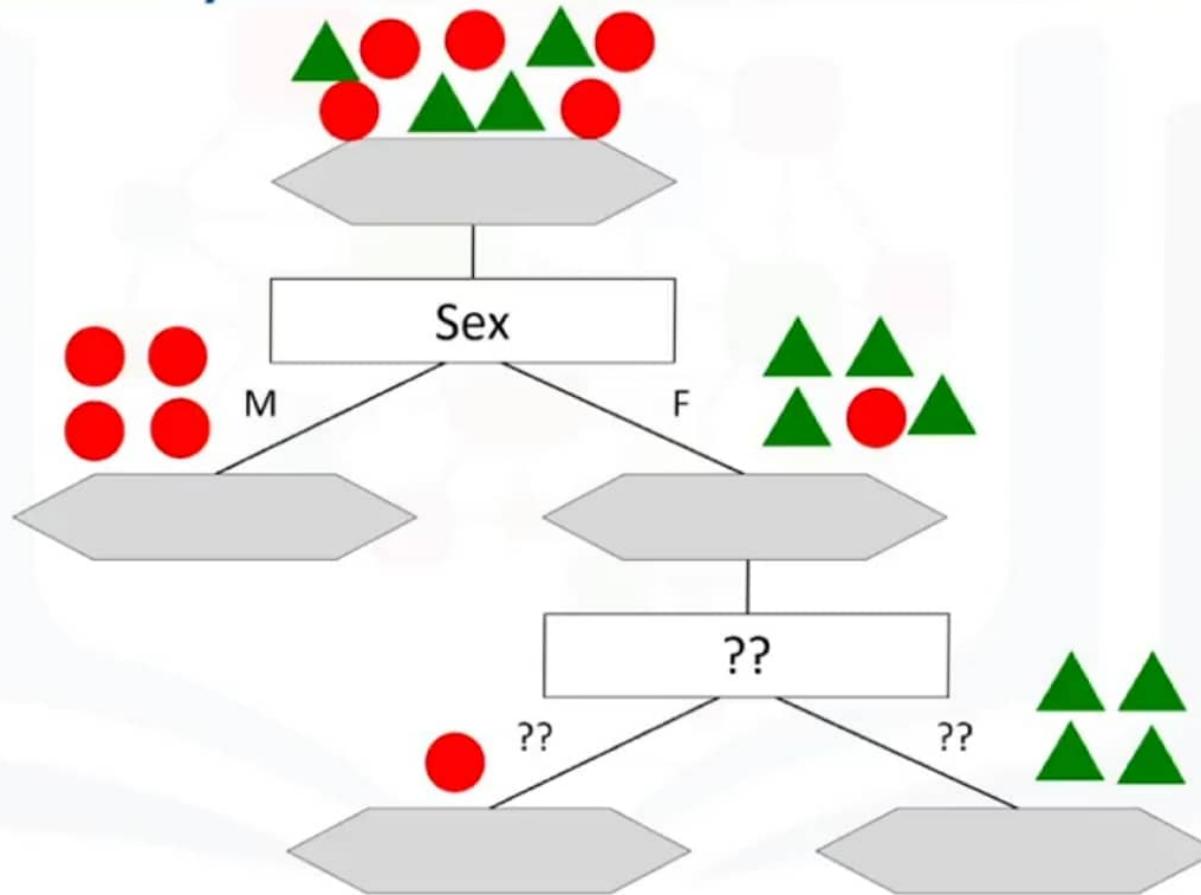


Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hhigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

?



Correct way to build a decision tree



What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.

	Independent variables										Dependent variable
	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn	
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes	
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes	
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No	
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No	
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?	

Continuous/Categorical variables

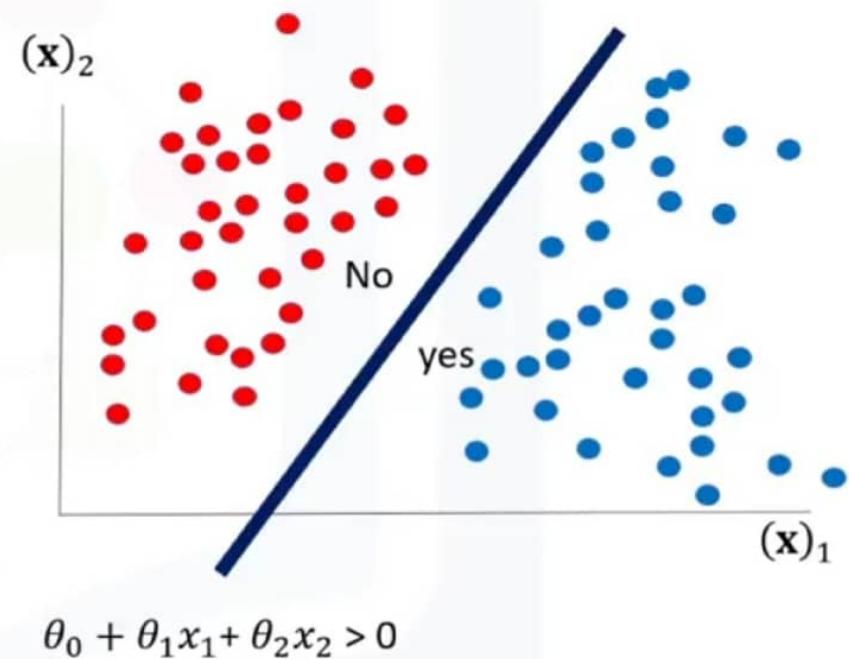
Categorical Variable

Logistic regression applications

- Predicting the probability of a person having a heart attack
- Predicting the mortality in injured patients
- Predicting a customer's propensity to purchase a product or halt a subscription
- Predicting the probability of failure of a given process or product
- Predicting the likelihood of a homeowner defaulting on a mortgage

When is logistic regression suitable?

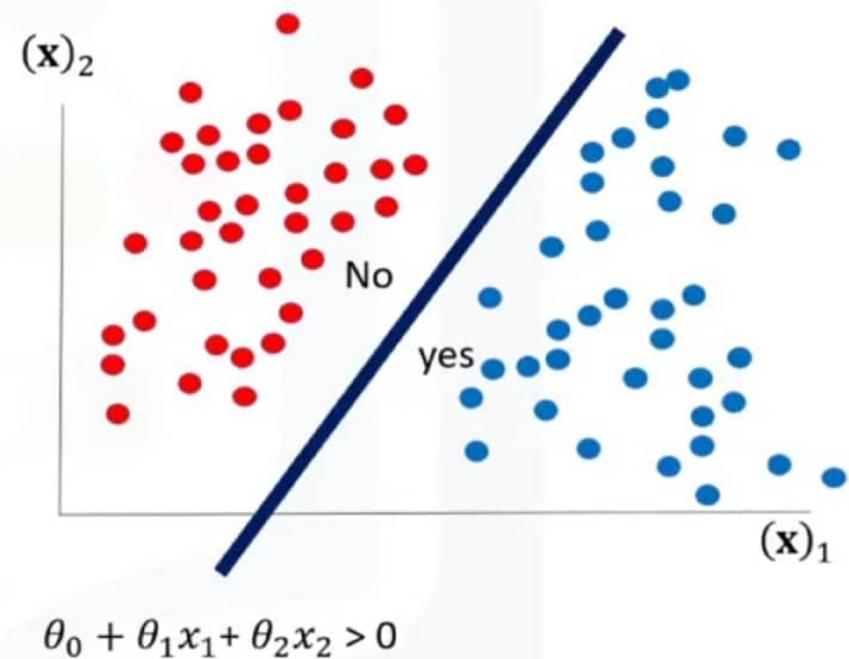
- If your data is binary
 - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary



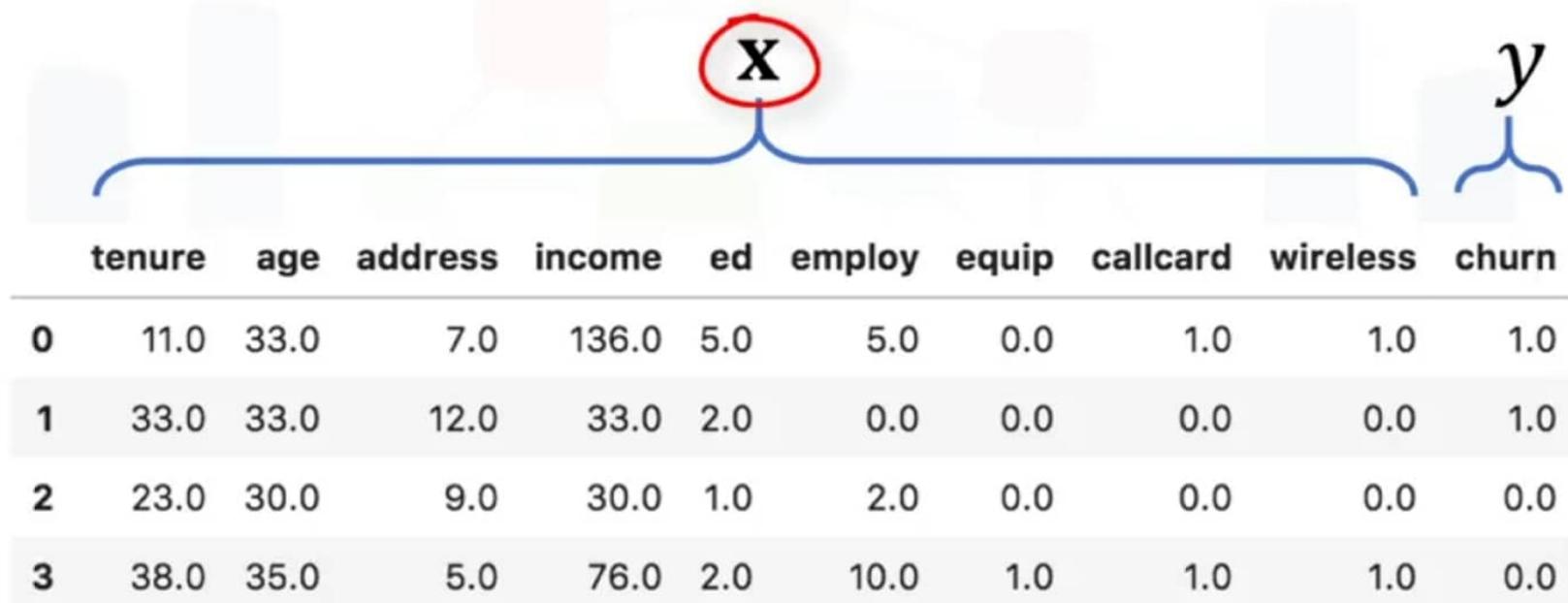
$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 > 0$$

When is logistic regression suitable?

- If your data is binary
 - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary
- If you need to understand the impact of a feature



Building a model for customer churn



The diagram illustrates a machine learning model structure. At the top, a red circle contains the letter 'X', positioned above a blue bracket that spans across the columns of a data table. Below this, another blue bracket points from the right side of the table towards the letter 'y'.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$$X \in \mathbb{R}^{m \times n}$$

$$y \in \{0,1\}$$

Building a model for customer churn

The diagram illustrates the inputs and output for a machine learning model. A horizontal bracket labeled 'X' covers the first eleven columns of the table, representing the input features: tenure, age, address, income, ed, employ, equip, callcard, wireless, and churn. A vertical bracket labeled 'y' covers the last column, representing the output target: churn.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$$X \in \mathbb{R}^{m \times n}$$

$$y \in \{0,1\}$$

$$\hat{y} = P(y=1|x)$$

$$P(y=0|x) = 1 - P(y=1|x)$$

Model of customer churn data

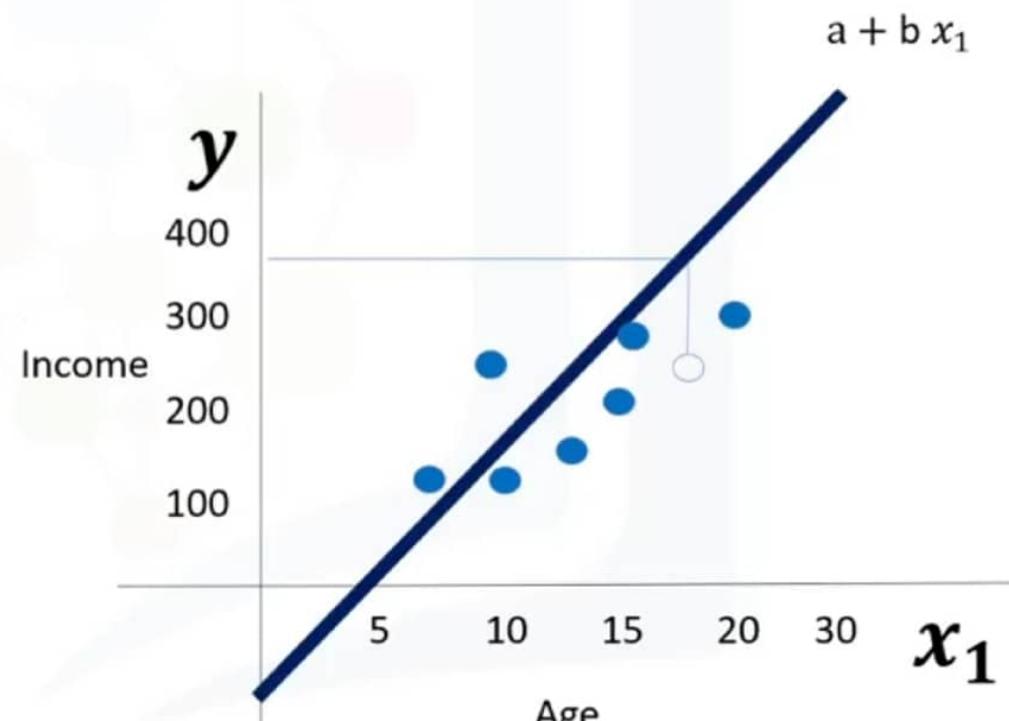
The diagram illustrates the structure of the customer churn data. A horizontal bracket labeled **x** covers all columns except the last one, which is labeled **y**. The columns represent various features: tenure, age, address, income, ed, employ, equip, callcard, wireless, and churn.

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

$$\hat{y} = P(y=1|x)$$

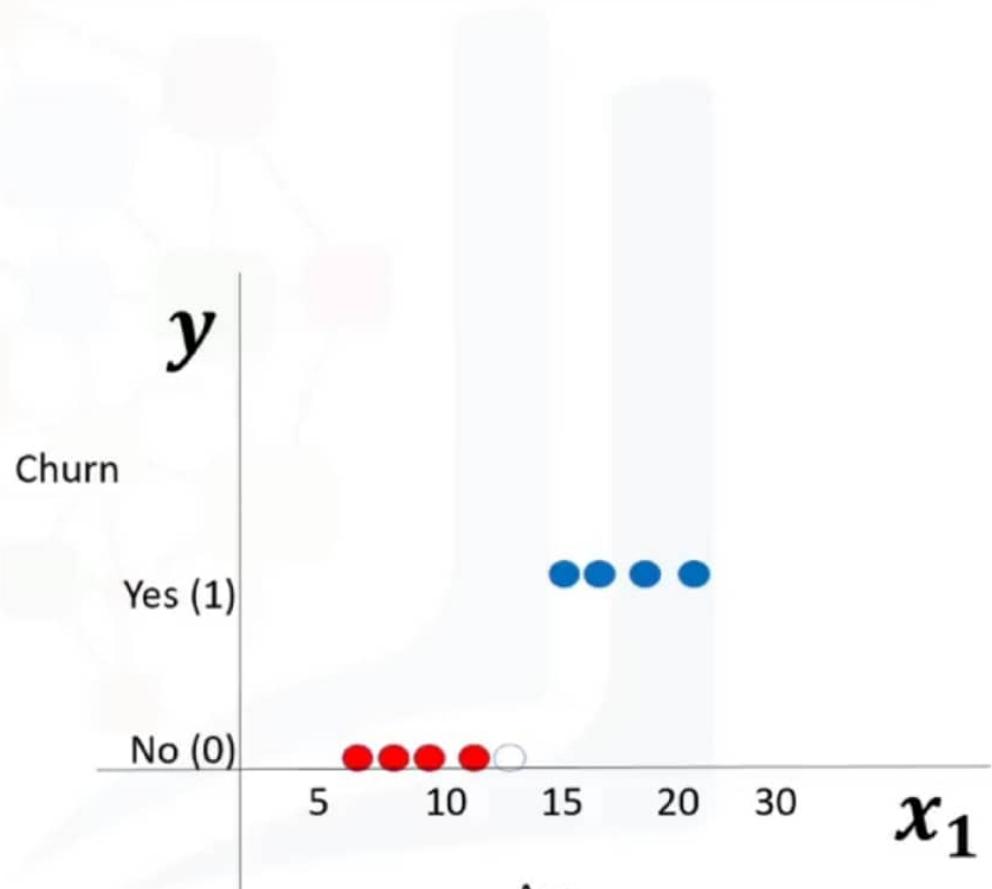
Predicting customer income

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Predicting churn using linear regression

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Predicting churn using linear regression

$$\theta^T = [\theta_0, \theta_1]$$

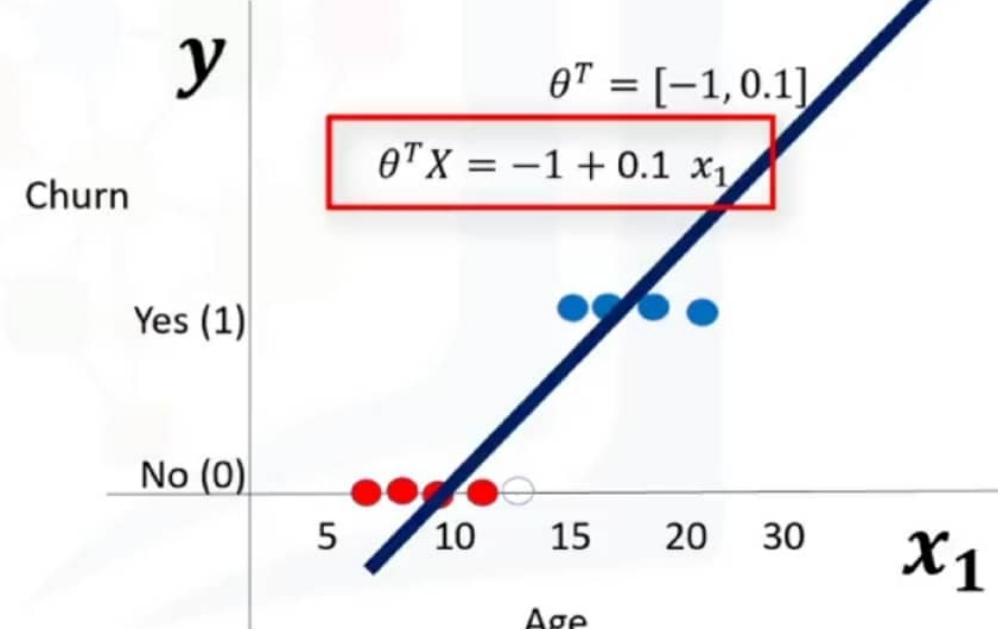
$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$\theta_0 + \theta_1 x_1$$

$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots$$

$$a + b x_1$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots] \quad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}$$



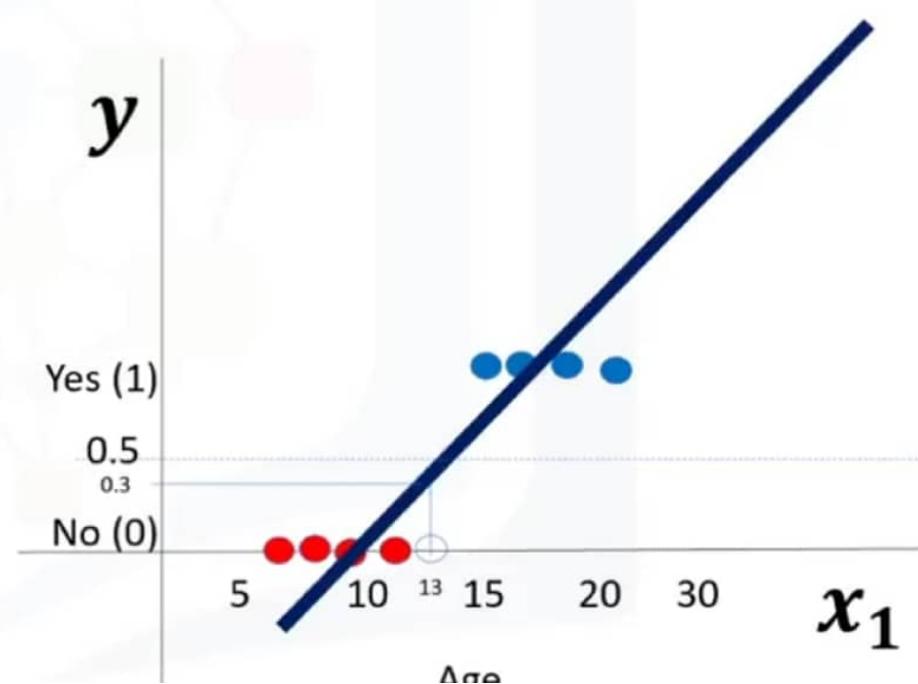
Linear regression in classification problems?

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$p_1 = [13] \rightarrow \theta^T X = -1 + 0.1 \cdot x_1 \\ = -1 + 0.1 \times 13 \\ = 0.3$$

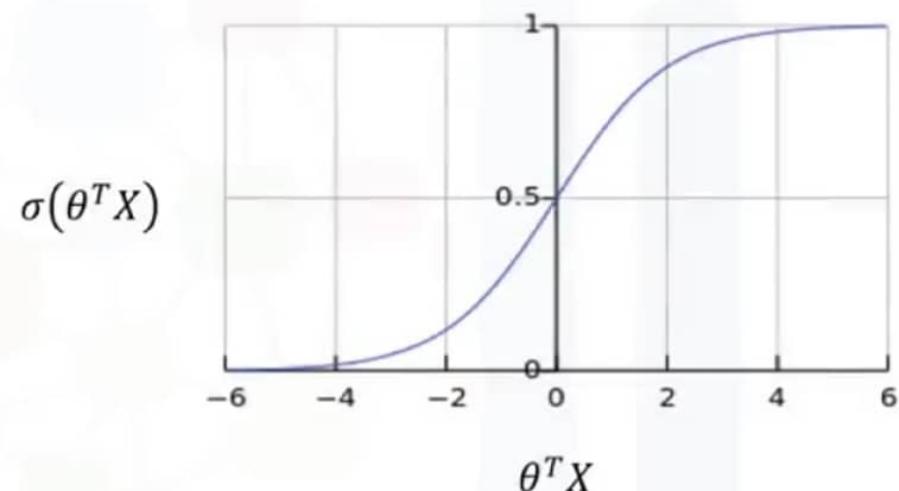
$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\theta^T X = -1 + 0.1 \cdot x$$



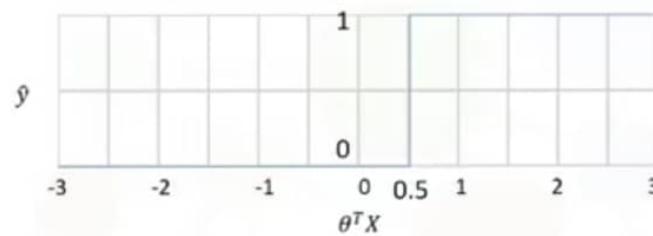
Sigmoid function in logistic regression

- Logistic Function



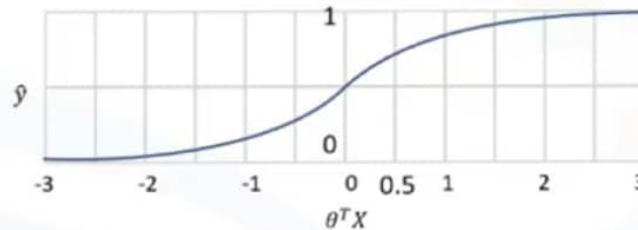
The problem with using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \dots$$



$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \dots)$$



$$\hat{y} = \sigma(\theta^T X)$$

$$P(y=1|x)$$

Sigmoid function in logistic regression

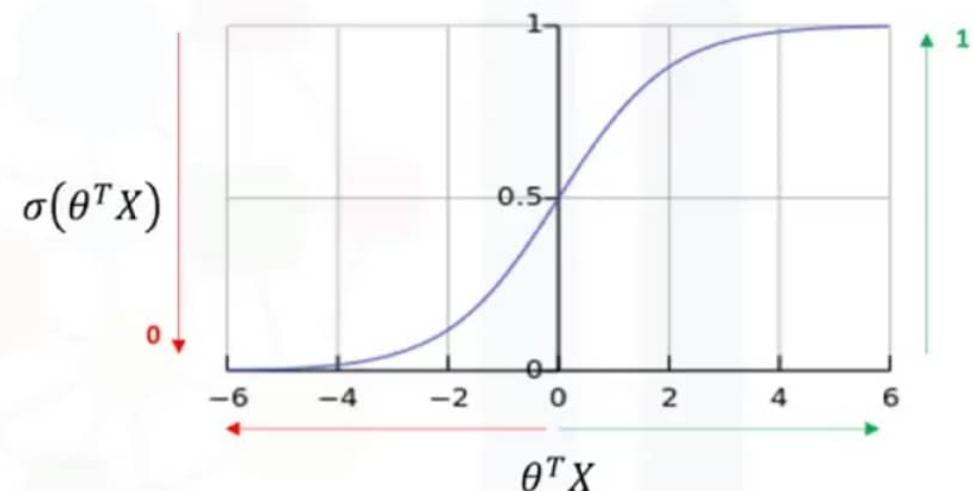
- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$$\sigma(\theta^T X) = 1$$

$$\sigma(\theta^T X) = 0$$

[0, 1]



Clarification of the customer churn model

What is the output of our model?

- $P(Y=1|X)$
 - $P(y=0|X) = 1 - P(y=1|x)$
-
- $P(\text{Churn}=1|\text{income,age}) = 0.8$
 - $P(\text{Churn}=0|\text{income,age}) = 1 - 0.8 = 0.2$

$$\sigma(\boldsymbol{\theta}^T \mathbf{X}) \longrightarrow P(y=1|x)$$

$$1 - \sigma(\boldsymbol{\theta}^T \mathbf{X}) \longrightarrow P(y=0|x)$$

The training process

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

1. Initialize θ .
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of \hat{y} with actual output of customer, y , and record it as error.
4. Calculate the error for all customers.
5. Change the θ to reduce the cost.
6. Go back to step 2.

$$\theta = [-1, 2]$$

$$\hat{y} = \sigma([-1, 2] \times [2, 5]) = 0.7$$

$$\text{Error} = 1 - 0.7 = 0.3$$

$$Cost = J(\theta)$$

$$\theta_{new}$$

General cost function

$$\sigma(\theta^T X) \longrightarrow P(y=1|x)$$

- Change the weight -> Reduce the cost

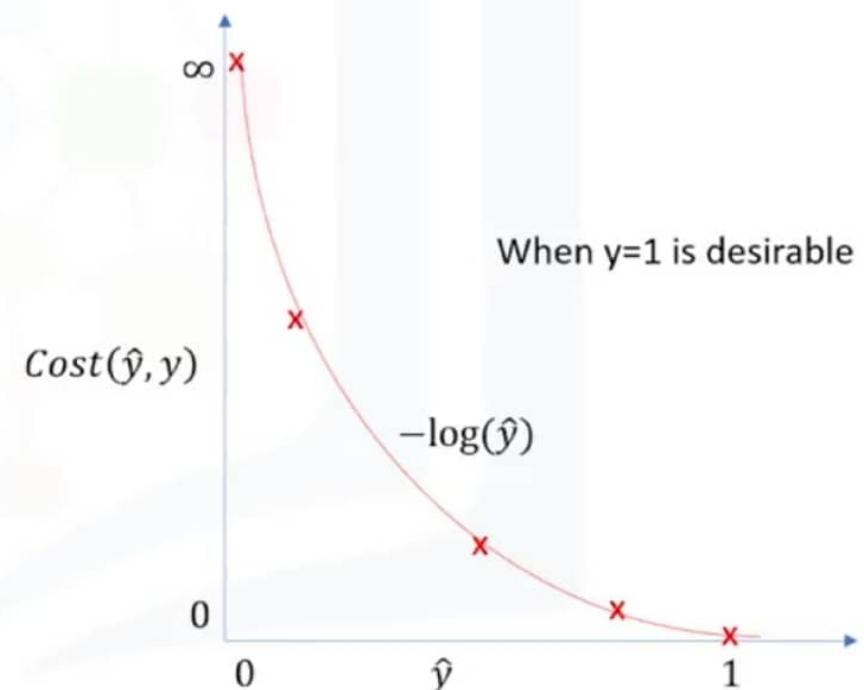
- Cost function

$$Cost(\hat{y}, y) = \frac{1}{2}(\sigma(\theta^T X) - y)^2$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(\hat{y}, y)$$

Plotting the cost function of the model

- Model \hat{y}
- Actual Value $y=1$ or 0
- If $Y=1$, and $\hat{y}=1 \rightarrow \text{cost} = 0$
- If $Y=1$, and $\hat{y}=0 \rightarrow \text{cost} = \text{large}$



Logistic regression cost function

- So, we will replace cost function with:

$$Cost(\hat{y}, y) = \frac{1}{2} (\sigma(\theta^T X) - y)^2$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m Cost(\hat{y}, y)$$

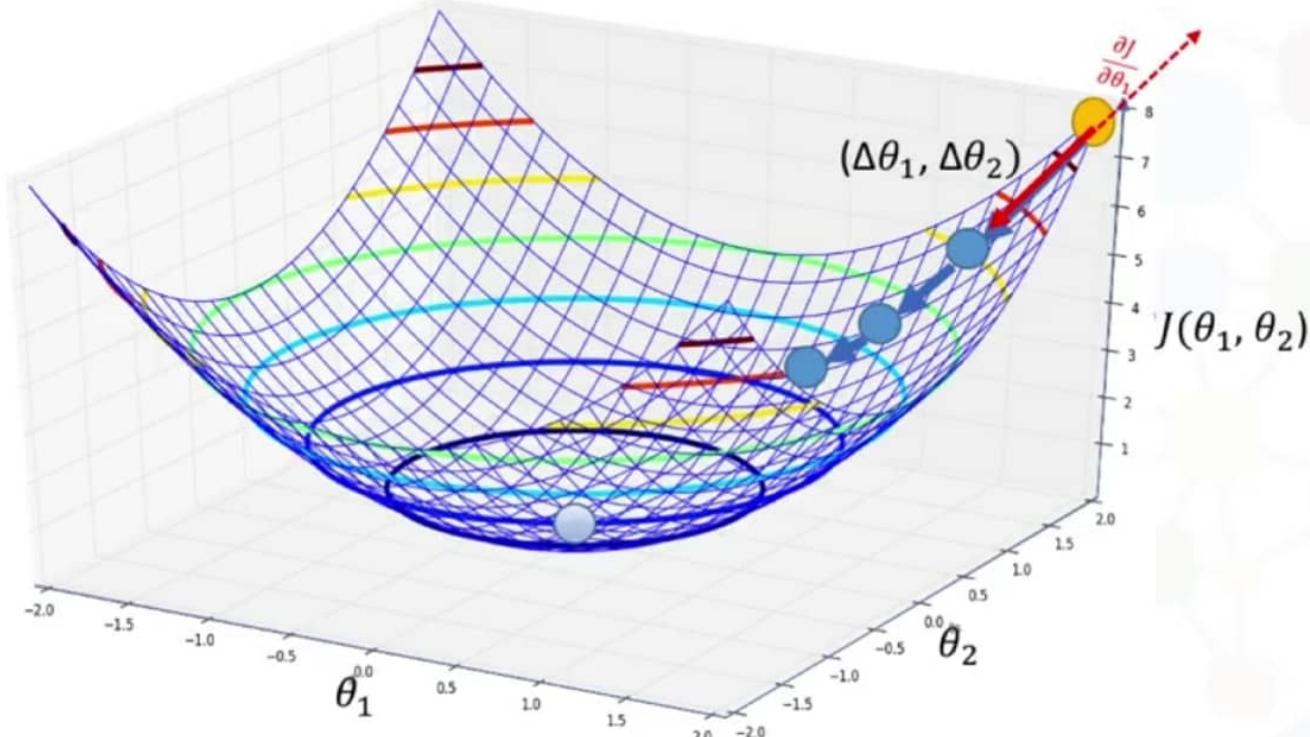
$$Cost(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1 - \hat{y}) & \text{if } y = 0 \end{cases}$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m y^i \log(\hat{y}^i) + (1 - y^i) \log(1 - \hat{y}^i)$$

Minimizing the cost function of the model

- How to find the best parameters for our model?
 - Minimize the cost function
- How to minimize the cost function?
 - Using Gradient Descent
- What is gradient descent?
 - A technique to use the derivative of a cost function to change the parameter values, in order to minimize the cost

Using gradient descent to minimize the cost



$$\hat{y} = \sigma(\theta_1 x_1 + \theta_2 x_2)$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m y^i \log(\hat{y}^i) + (1 - y^i) \log(1 - \hat{y}^i)$$

$$\frac{\partial J}{\partial \theta_1} = -\frac{1}{m} \sum_{i=1}^m (y^i - \hat{y}^i) x_1^i$$
$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \frac{\partial J}{\partial \theta_3} \\ \vdots \\ \frac{\partial J}{\partial \theta_k} \end{bmatrix}$$

$$\text{New } \theta = \text{old } \theta - \eta \nabla J$$

Training algorithm recap

1. initialize the parameters randomly.
2. Feed the cost function with training set, and calculate the error.
3. Calculate the gradient of cost function.
4. Update weights with new values.
5. Go to step 2 until cost is small enough.
6. Predict the new customer X.

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots]$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m y^i \log(\hat{y}^i) + (1 - y^i) \log(1 - \hat{y}^i)$$

$$\nabla J = \left[\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \frac{\partial J}{\partial \theta_3}, \dots, \frac{\partial J}{\partial \theta_k} \right]$$

$$\theta_{new} = \theta_{prev} - \eta \nabla J$$

$$P(y=1|x) = \sigma(\theta^T X)$$

Classification with SVM

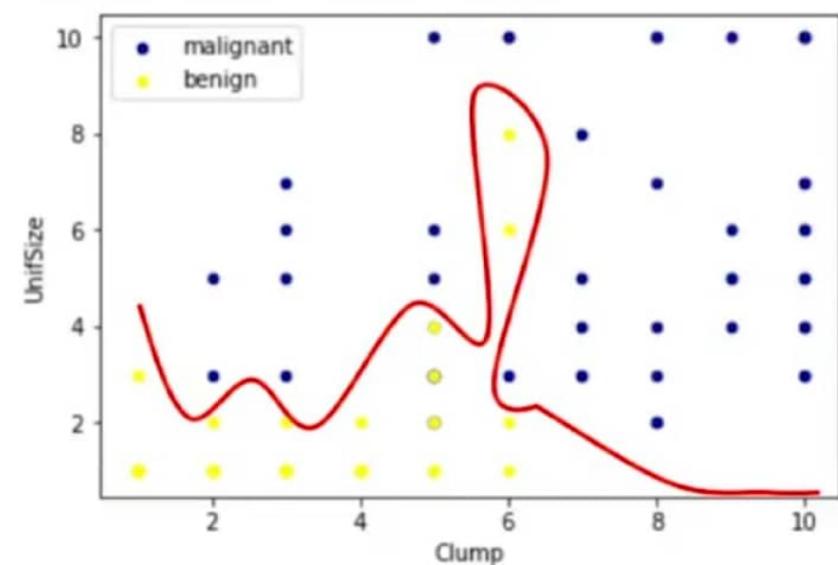


What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

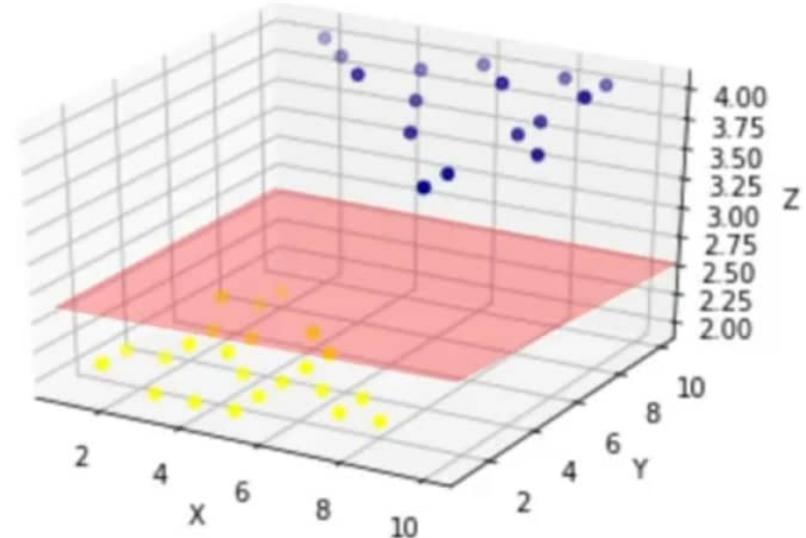


What is SVM?

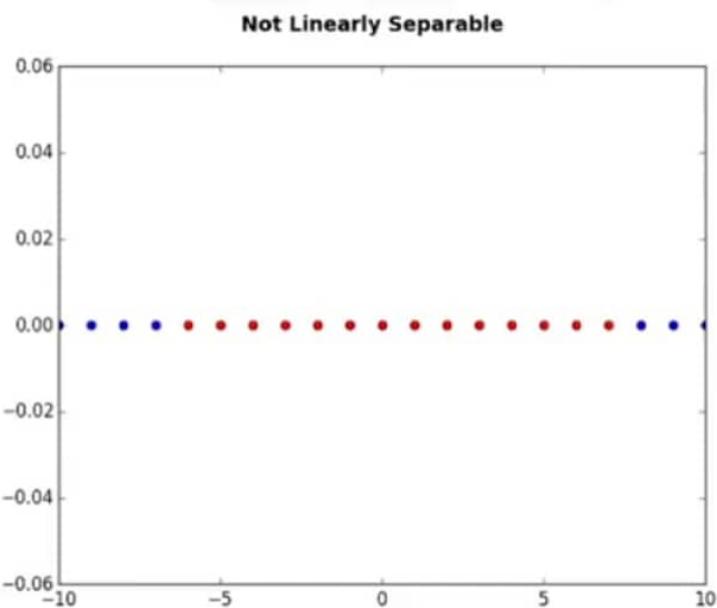
SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

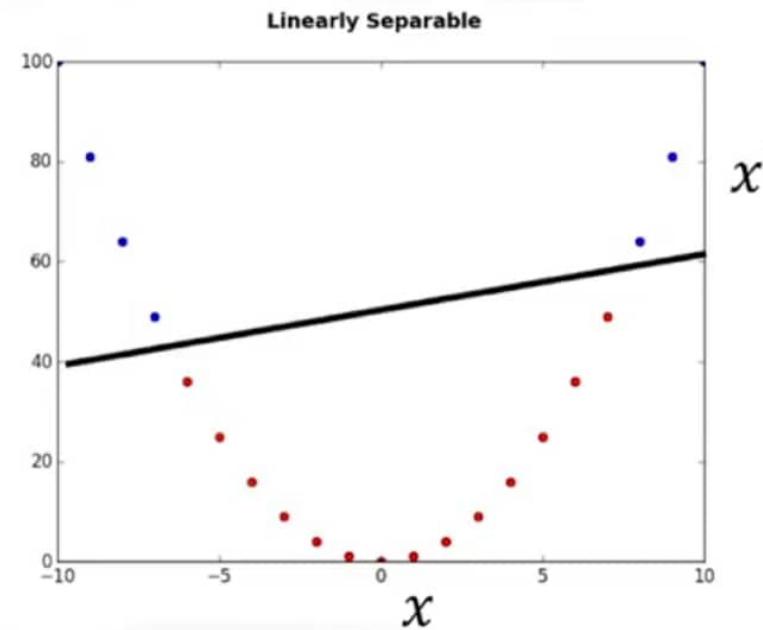
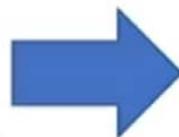
Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign



Data transformation



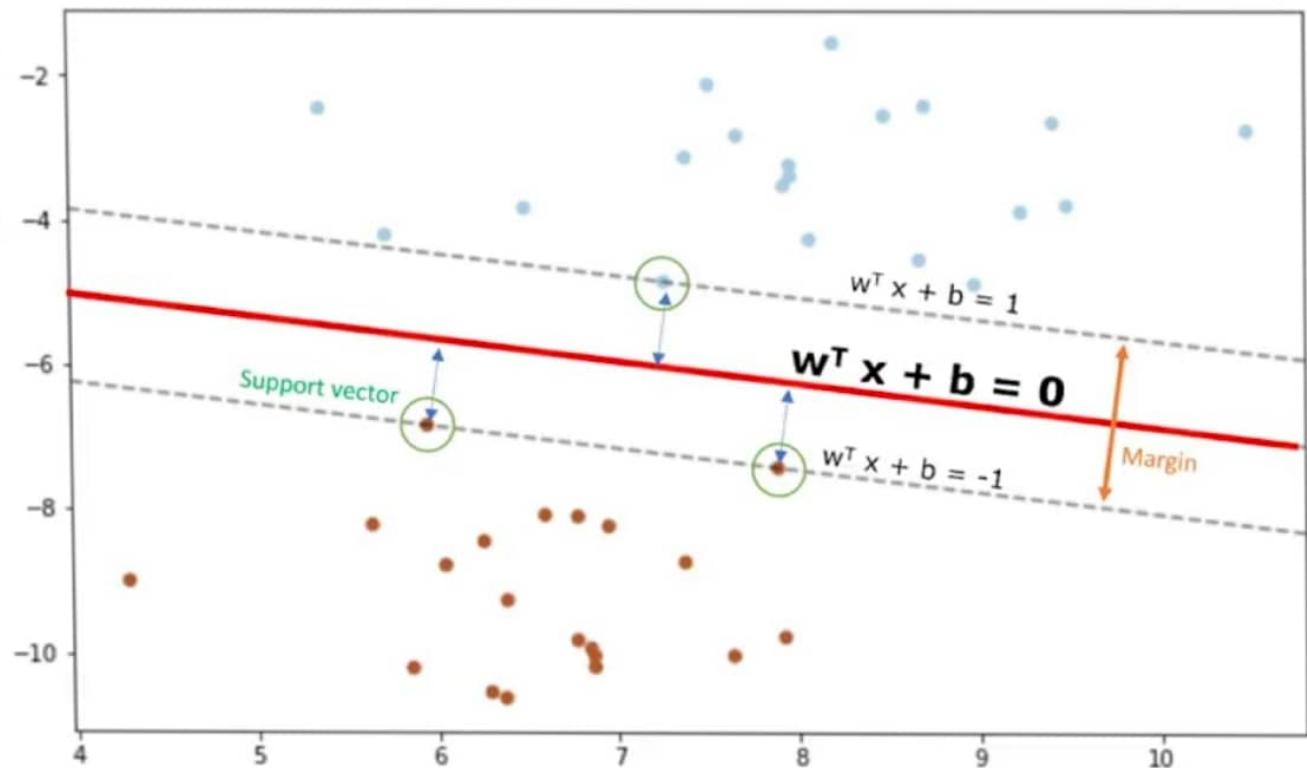
- Kernelling:
- Linear
 - Polynomial
 - RBF
 - Sigmoid



$$\phi(x) = [x, x^2]$$

Using SVM to find the hyperplane

Find \mathbf{w} and b such that
 $\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{w}$ is minimized;
and for all $\{(\mathbf{x}_i, y_i)\}$: $y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1$



Pros and cons of SVM

- Advantages:

- Accurate in high-dimensional spaces
- Memory efficient

- Disadvantages:

- Prone to over-fitting
- No probability estimation
- Small datasets

SVM applications

- Image recognition
- Text category assignment
- Detecting spam
- Sentiment analysis
- Gene Expression Classification
- Regression, outlier detection and clustering